GUSTO Data Vault: Laying the foundations for an open science system with OMOP Data Catalogue

Cindy Ho[1,2], Li Ting Ang[1,2], Maisie Ng[1,2], Hang Png[1,2], Shuen Lin Tan[1,2], Estella Ye[1,2], Sunil Kumar Raja[1], Mengling Feng[3,4], Johan G Eriksson[1,5,6,7], Mukkesh Kumar[1,2,3]

Institution(s) of origin:

[1] Singapore Institute for Clinical Sciences, Agency for Science Technology and Research, Singapore
[2] Bioinformatics Institute, Agency for Science Technology and Research, Singapore
[3] Saw Swee Hock School of Public Health, National University of Singapore, National University Health System, Singapore
[4] Institute of Data Science, National University of Singapore, Singapore
[5] Department of Obstetrics and Gynaecology and Human Potential Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6] Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland
[7] Folkhälsan Research Center, Helsinki, Finland

Background

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) created by Observational Health Data Sciences and Informatics (OHDSI), is an open community data standard, designed to standardise the structure and content of observational data[1]. The harmonisation of data using OMOP CDM enables systematic and collaborative OMOP-based research. The Agency for Science, Technology and Research (A*STAR) has been managing the datasets for national level research programmes such as the Growing Up in Singapore Towards healthy Outcomes (GUSTO)[2] birth cohort study. A*STAR has pioneered health research with an open interactive GUSTO Data Vault platform (Figure 1) for hypothesis construction and data-driven discoveries[3,4]. The OMOP Data Catalogue in GUSTO Data Vault showcases the data assets which have been converted into Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) format. The current CDM version is CDM v5.4[5], as depicted below for GUSTO tables and fields (Figure 2).

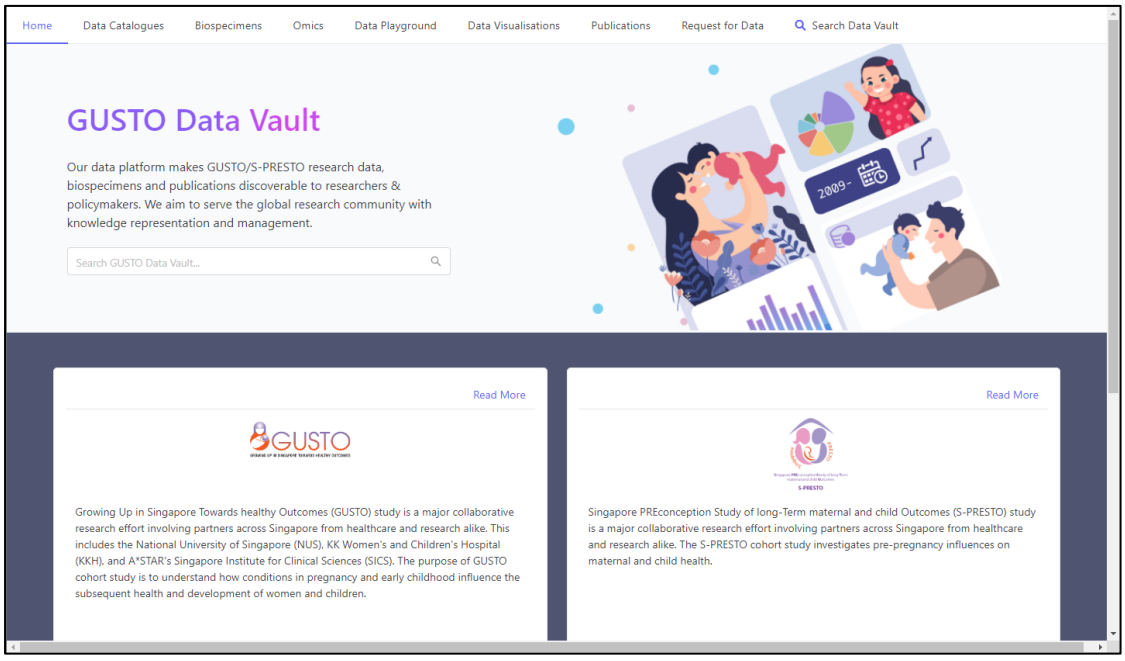

Figure 1. GUSTO Data Vault Homepage

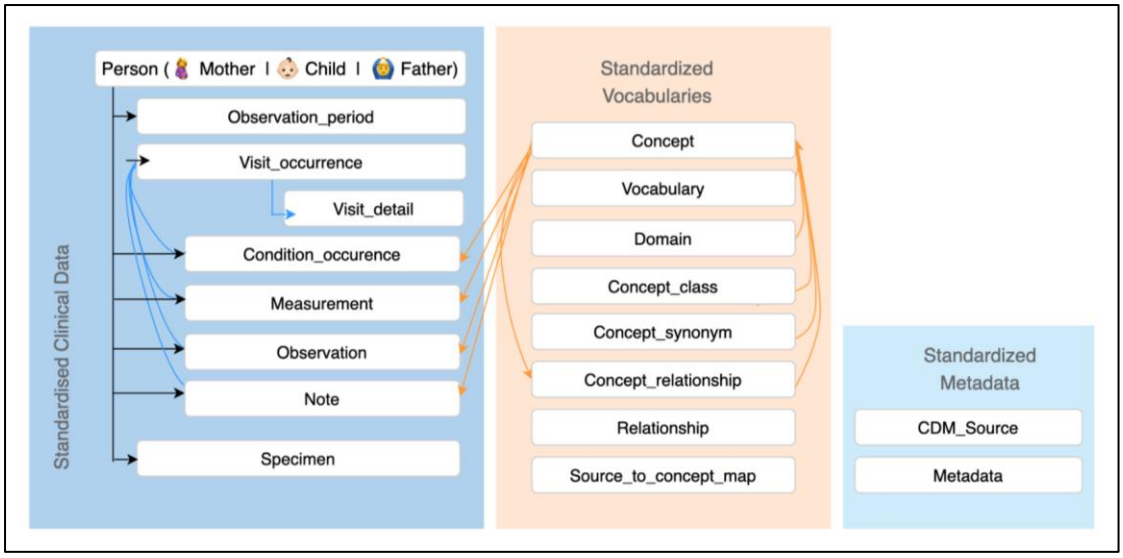

Figure 2. OMOP CDM Model for GUSTO

Methods

Reinventing the working culture around OMOP data literacy, research organisations and pharmaceutical companies can create a data governance structure with OMOP Data Catalogue. The OMOP Data Catalogue was developed using open-source technologies. The containerised web application with Docker was built using PostgreSQL database engine and Django web application framework. The GUSTO Data Vault application is deployed on AWS cloud environment. The front-end (client side) scripting is using HTML, CSS, jQuery, Ajax. The back-end (server side) scripting is using Python. The GUSTO OMOP fields showcased in OMOP Data Catalogue were mapped using open source OMOP tools provided by OHDSI and R programming community.

Results

The FAIR (Findable, Accessible, Interoperable, Reusable) guiding principles for data management and stewardship is a critical need for academia, industry, funding agencies and scholarly publishers[6]. The OMOP Data Catalogue makes GUSTO cohort-specific CDM fields to be discovered by the global research community. The OMOP CDM fields can be discovered across the Person, Condition, Observation and Measurement tables (Figure 3). Researchers can perform search across variables and concepts. The metadata is described with a plurality of relevant attributes such as CDM Field, Concept ID, Concept Type, Subject Type, Visit Timepoint, Description and Domain. The data profiling of the OMOP Concept IDs enables GUSTO data to be identified, described, discovered, and reused by researchers (FAIR data principles). The visual interface of OMOP Data Catalogue guides researchers through the GUSTO tables and fields. The OMOPed data from incremental OMOP conversions (e.g. Table-by-table conversions) can be seamlessly integrated in OMOP Data Catalogue by GUSTO data curators.

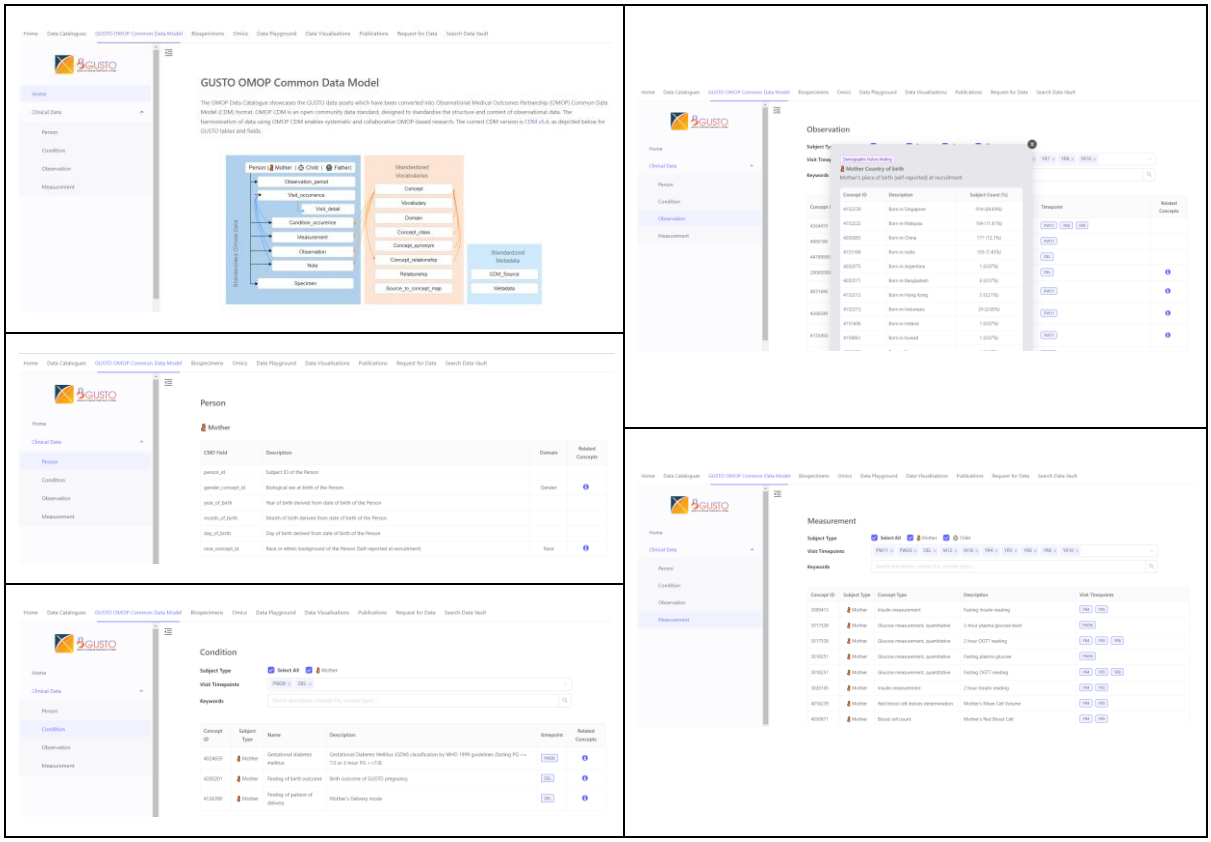

Figure 3. OMOP Tables for GUSTO

Conclusions

The database schemas can vary for population health studies and involves considerable data engineering efforts for data standardisation. Singapore have initiated national level OMOP mapping to curate clinical datasets in OMOP standard to improve data interoperability. The GUSTO OMOP Data Catalogue developed for a Singapore cohort study lays the foundations for developing cross-study OMOP Data Catalogues expanded across APAC and global OHDSI data partners, enabling database level characterizations for knowledge discovery and management in an open science system.

References/Citations

1. Observational Health Data Sciences and Informatics. The Book of OHDSI [Internet]. [place unknown: Observational Health Data Sciences and Informatics]; 2021 [cited 2023 Apr 10] Available from: https://ohdsi.github.io/TheBookOfOhdsi/.
2. Soh S-E, Tint MT, Gluckman PD, Godfrey KM, Rifkin-Graboi A, Chan YH et al. Cohort profile: Growing Up in Singapore Towards healthy Outcomes (GUSTO) birth cohort study. Int J Epidemiol. 2014 Oct;43(5):1401-9. DOI: 10.1093/ije/dyt125.
3. Agency for Science, Technology and Research. GUSTO Data vault [Internet]. Singapore: Agency for Science, Technology and Research; [cited 2023 Apr 10]. Available from: https://gustodatavault.sg/.
4. Ho C, Ang LT, Ng M, Png H, Tan SL, Ye E, et al. GUSTO Data Vault: Working Towards OMOP Data Standardisation. Poster session presented at: 2022 APAC OHDSI Symposium; 2022 Nov 13; Taiwan.
5. Observational Health Data Sciences and Informatics. OMOP Common Data Model [Internet]. [place unknown: Observational Health Data Sciences and Informatics]; [cited 2023 Apr 10]. Available from: https://ohdsi.github.io/CommonDataModel/cdm54.html.
6. Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3 [Internet]. 2016 Mar [cited 2023 Apr 10]; 160018 (2016). Available from: https://www.nature.com/articles/sdata201618 DOI: 10.1038/sdata.2016.18